CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

DSM

IN2P3 INSTITUT NATIONAL DE PHYSIQUE NUCLÉAIRE ET DE PHYSIQUE DES PARTICULES

LCG-France Tier-1 and Analysis Facility

Overview

Fabio Hernandez
IN2P3/CNRS Computing Centre - Lyon
fabio@in2p3.fr

Atlas Tier-1 tour
Lyon, April 26th-27th 2007

Contents

- Resources
  - Budget, plan and pledges
  - contribution
- Current and future work
- Conclusions
- Questions

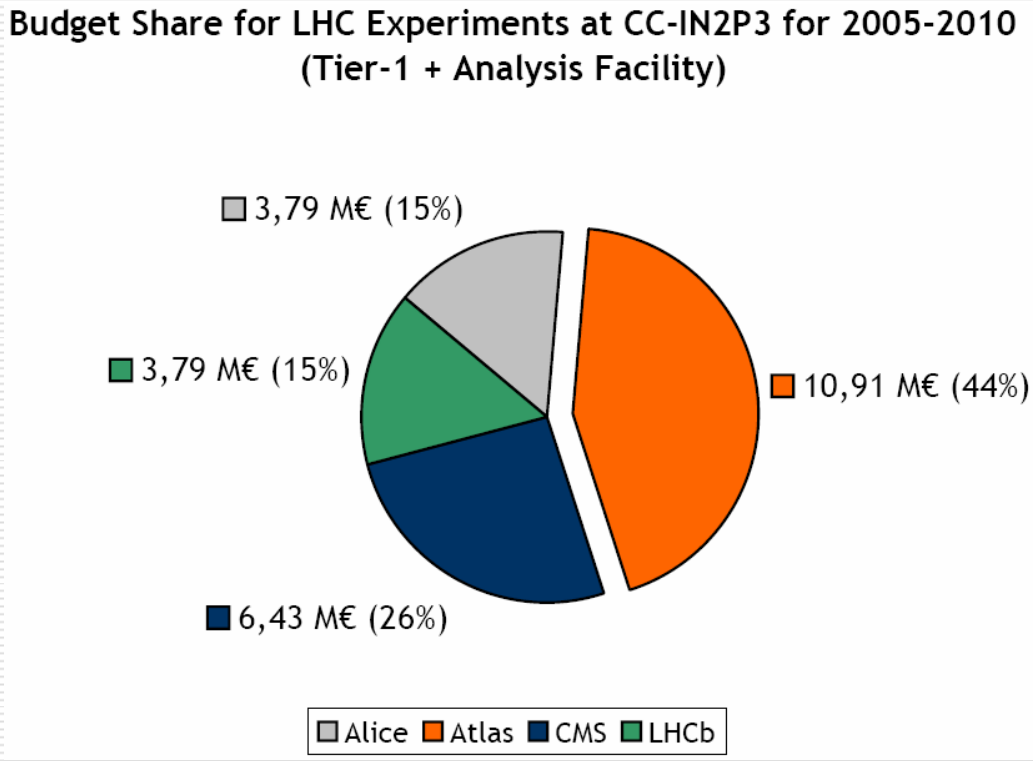
Overview

- Shared facility
  - Tier-1 and Analysis facility for all the LHC experiments
  - Several experiments, including DZero, Babar, Virgo in the top 10 CPU consumers
- Also operating grid services for non-LHC VOs

Grid Service	alice	atlas	cms	lhcb	auvergrid	biomed	calice	cdf	dteam	dzero	egeode	embrace	esr	hone	ilc	ops	virgo
CE	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓
dCache/SRM SE	✓	✓	✓	✓					✓							✓	
Classic SE	✓	✓	✓	✓		✓			✓	✓	✓		✓	✓	✓	✓	✓
Local LFC	✓	✓	✓	✓													
VO Box	✓	✓	✓	✓				✓									
FTS	✓	✓	✓	✓													
Central LFC						✓											
RLS/RMC						✓											
VOMS					✓	✓					✓	✓					

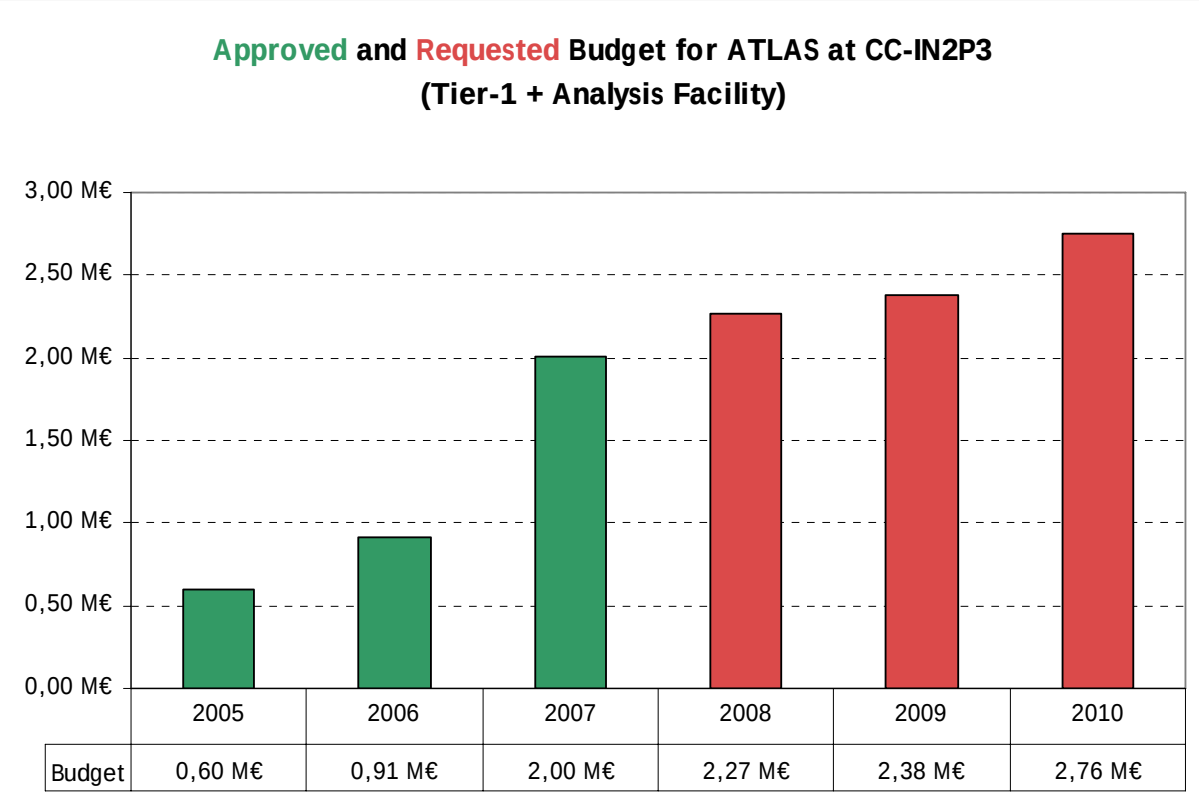
Budget: all LHC Experiments

- Equipment and running costs for all LHC experiments at CC-IN2P3 (2005-2010)
  - Total required: 24,9 M€
    - the refurbishment of current machine room and the construction of a second one are NOT included
- Budget requested on a pluriannual basis
  - Approval on a yearly basis
  - Impact on hardware procurement process



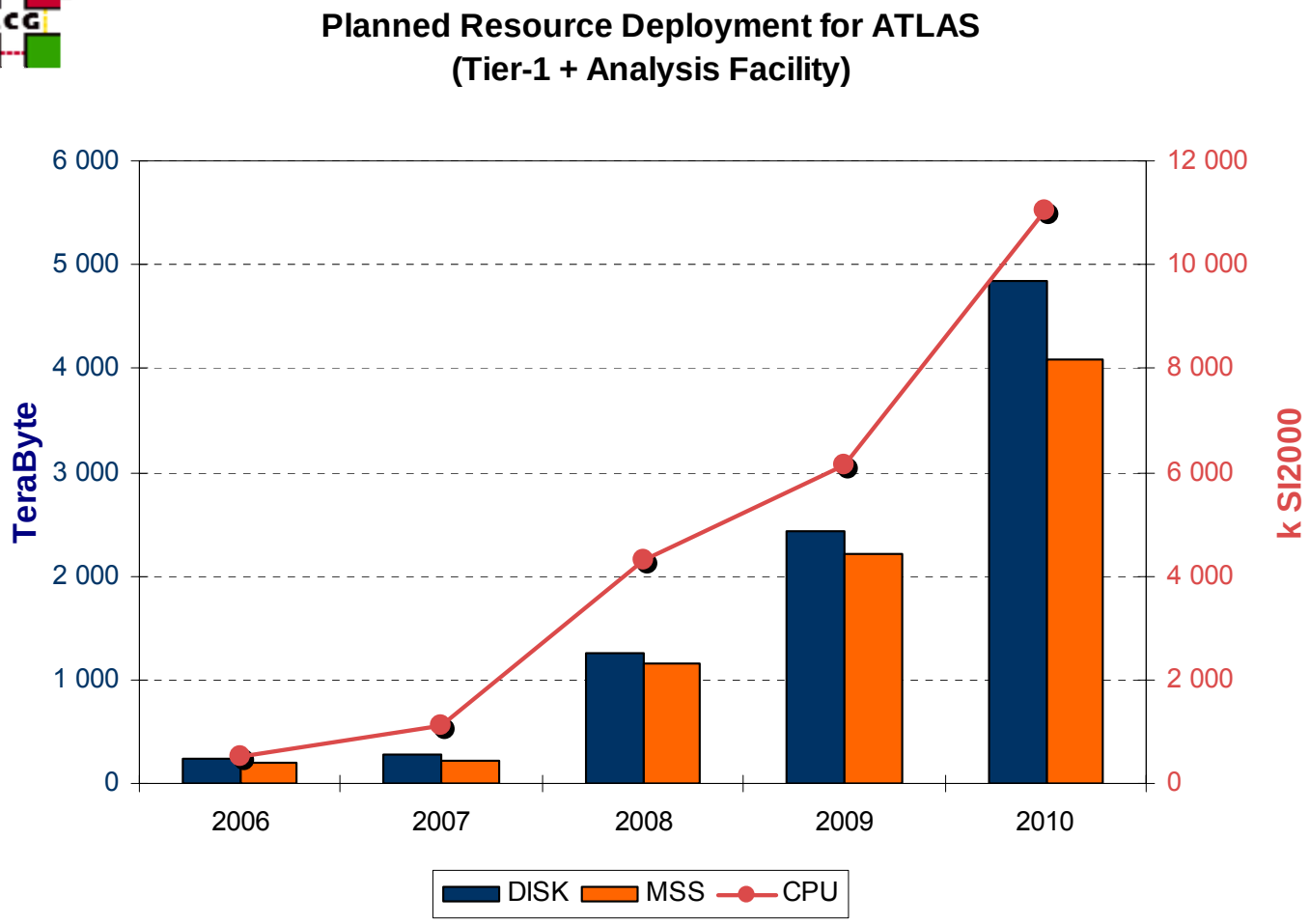
Budget: Atlas

- Equipment and running costs for Atlas needs (2005-2010)
  - 10,9 M€



	2005	2006	2007	2008	2009	2010
Budget	0,60 M€	0,91 M€	2,00 M€	2,27 M€	2,38 M€	2,76 M€

Planned Resource Deployment



Pledges

LCG-France Tier-1 at CC-IN2P3: Pledged Resources

		2007	2008	2009	2010
Atlas	CPU [k SI2000]	362	2 066	3 240	5 651
	Disk [TB]	246	1 133	2 244	4 502
	MSS [TB]	176	877	1 704	3 272

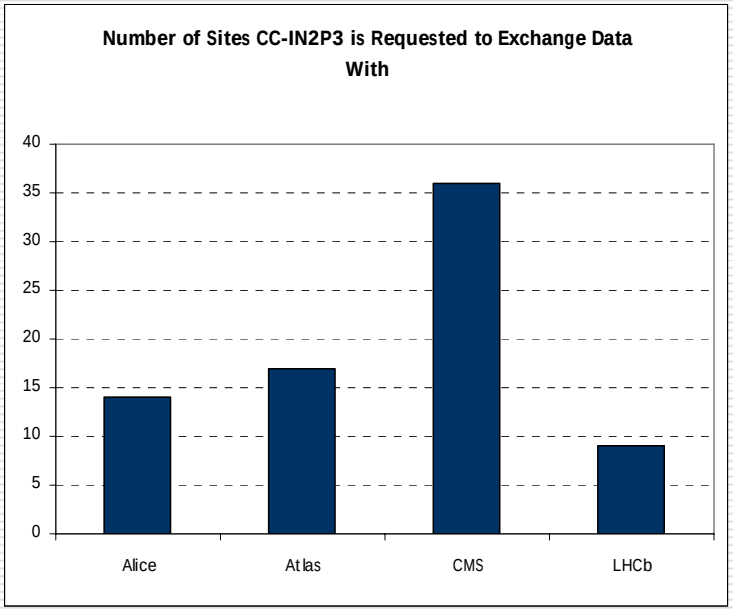
LCG-France Analysis Facility at CC-IN2P3: Pledged Resources

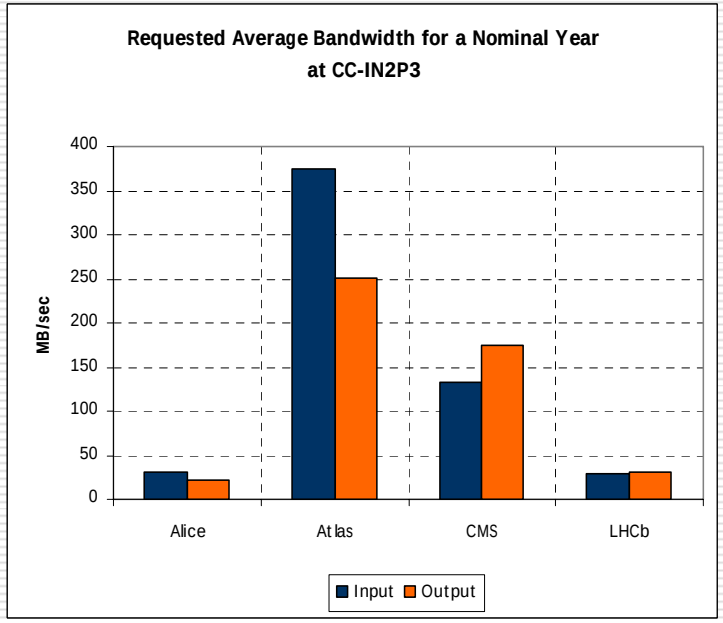
		2007	2008	2009	2010
Atlas	CPU [k SI2000]	78	583	899	1 718
	Disk [TB]	3	12	18	33
	MSS [TB]	0	0	0	0

- Note: a fraction of the resources for the Analysis Facility are reserved for ATLAS-France usage, so they are not pledged

WAN bandwidth requirements

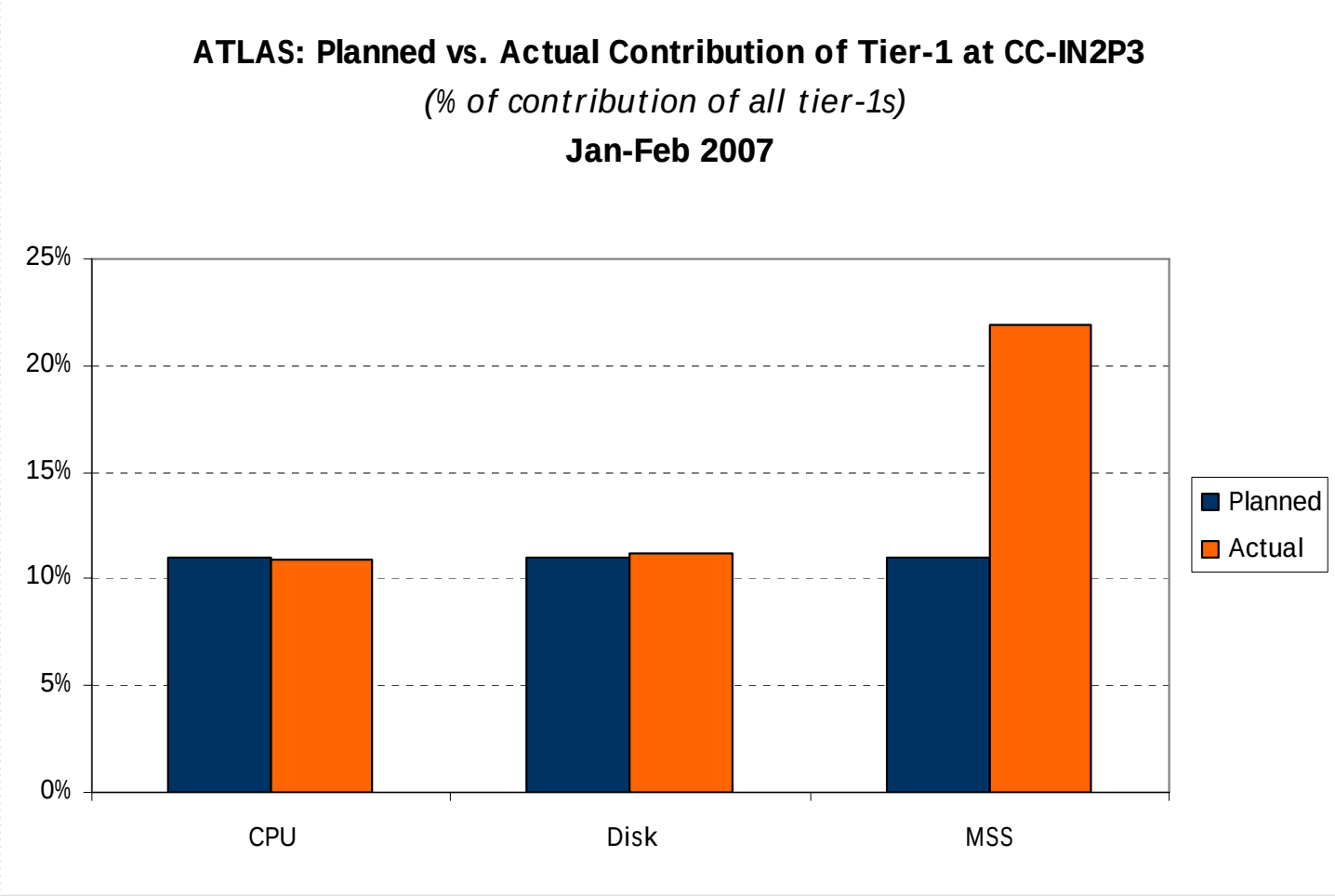
Experiment	Number of Sites	Input		Output	
		Average Bandwidth [MB/sec]	Peak Bandwidth [MB/sec]	Average Bandwidth [MB/sec]	Peak Bandwidth [MB/sec]
Alice	14	30,7	40,7	22,3	29,5
Atlas	17	373,8	522,4	251,6	359,8
CMS	36	132,7	132,7	174,2	404,2
LHCb	9	28,4	28,4	31,8	31,8
	Total	565,6	724,2	479,9	825,3





Source: Megatable http://lcg.web.cern.ch/LCG/documents/Megatable240107.xls

Planned vs. Actual Contribution



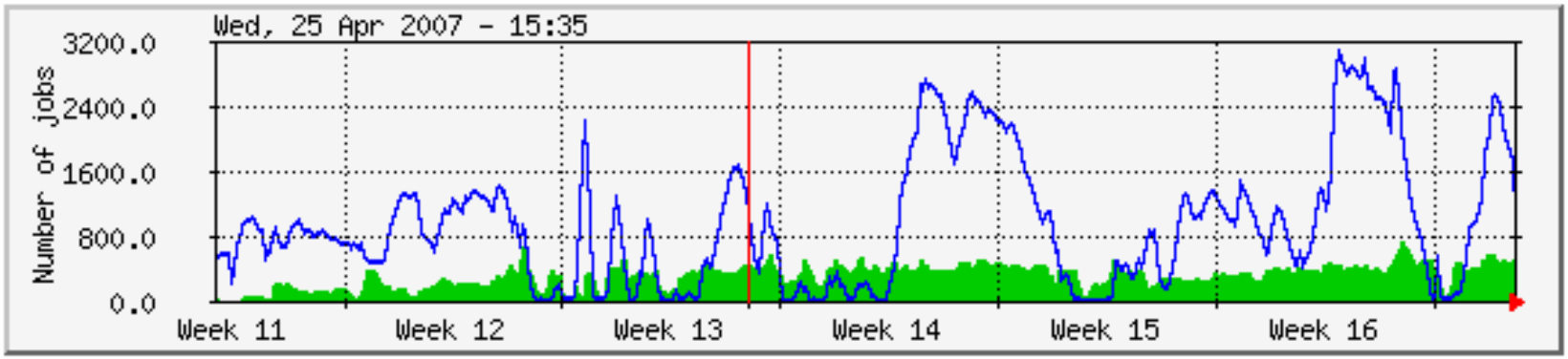
Source: http://lcg.web.cern.ch/LCG/MB/accounting/accounting_summaries.pdf

Planned vs. Actual Contribution: CPU

- Available capacity for Atlas
  - 2007H1: 500 kSI2000
    - Roughly 416 cores (AMD Opteron 275, 2.2 GHz)
  - 2007H2: 1097 kSI200 (tier-1 + AF)
    - Only 440 kSI2000 are pledged

Monthly graph (2 hours average)

Atlas Jobs Queued and Running



Max Jobs running: 749 jobs. Average Jobs running: 324 jobs. Current Jobs running: 534 jobs.

Max Jobs in queue: 3085 jobs. Average Jobs in queue: 958 jobs. Current Jobs in queue: 981 jobs.

Source: http://cctools.in2p3.fr/mrtguser/mrtguser/atlas/atlas_jobs.html

Planned vs. Actual Contribution: Disk

- Capacity for Atlas
  - 2007H1
    - Pledged: 240 TB
    - Delivered: 149 TB
      - *22 TB free (as of 25/04/2007)*
  - 2007H2
    - Pledged: 249 TB
    - Total capacity to be delivered: 271 TB (tier-1 + AF)

Planned vs. Actual Contribution: MSS

- Capacity for Atlas
  - 2007H1
    - Pledged: 200 TB
    - Used: 186 TB
  - 2007H2
    - Pledged: 176 TB (tier-1 + AF)
      - *All the pledged MSS capacity for 2007 has already been consumed*
    - Total capacity to be delivered: 222 TB

Resource Deployment

- According to Harry Renshall's table, for 2007Q2 Atlas requires our site to provide
  - Disk: 52 TB
    - to be compared to the current allocation of 149 TB
  - MSS: 91 TB
    - To be compared to the used space 186 TB
  - Source: https://twiki.cern.ch/twiki/bin/view/LCG/SC4ExperimentPlans

- A realistic plan of when the pledged resources are really needed in our site is highly desirable

Procurement

- We have to satisfy several constraints
  - Formal process is long
    - Call for tenders at the European level
  - Budget is approved on a yearly basis
    - « Final word » during last quarter each year
  - Limited machine room space available
    - Extensive in-situ tests to realistically identify the real caracteristics of candidate hardware (when possible)
    - Optimize (computing power/m²) but also (computing power/€)
      - *Forecast of running costs performed at this stage*
  - Desired availability of computing equipment in operation by experiments
  - Very frequent delays in the deliveries of equipment

Procurement (cont.)

- Starting in 2007, we are modifying our procurement plan for LHC experiments
  - With budget for year N, purchase 40% of the required equipment for year N+1
  - Procurement process for remaining 60% triggered as soon as next year budget is known
- Consequence
  - Hardware purchased in advance potentially more expensive, so less computing capacity for the same money
    - This is the reason of the decrease of our revised pledges with respect to the previous plan…
    - …in spite of the fact that our planned requested budget was not decreased
  - Special case for 2007
    - We will purchase all the hardware required for year 2007 plus 40% of our pledges for 2008
- If this model works well, we would be able to provide a significant amount of the pledged capacity by April 1st each year
  - Again, a detailed schedule of the expected usage by the experiments would be very helpful
    - This won't change the procurement but will help us with the deployment
- Providing 100% of the 2008 pledged resources on January 2008, as recently requested by Atlas, is simply not possible for us

2007: Compute Capacity Increase

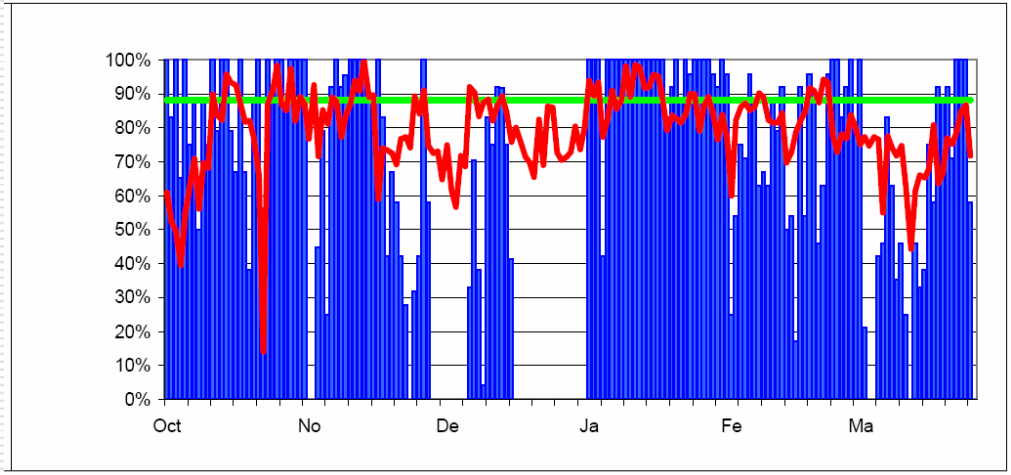
- New cartridge library being commissioned
  - SUN/STK SL8500, 30 drives, 10.000 slots, up to 5 PB
- On-going call for tenders for compute nodes and disk servers
  - +4,5 M SI2000
    - Non-LHC: 1 M SI2000
    - LHC
      - *Needs for 2007: 1,3 M SI2000*
      - *Provision for 2008: 2,2 M SI2000 (~40% of capacity required in 2008)*
  - +1200 TB (DAS)
    - LHC needs for 2007: 400 TB
    - LHC provision for 2008: 800 TB
  - Expected (contractual) delivery early July
- Additional tender for +160 TB (SAN)
  - GPFS, somes dCache spaces, HPSS disk cache

Facility Upgrade

- Major effort for upgrading the electric and cooling infrastructure of the site
  - Budget: more than 1,5 M€
  - The limits of the current capacity was reached this quarter
    - Severe power outage early March
    - Since then, a fraction of the worker nodes are out of the UPS circuitry
  - From 500 kW to 1000 kW of electrical power usable for computing equipment
    - +600 kW for cooling
    - Expected availability of this increased capacity: June 2007
  - New diesel generator
  - Additional UPS
  - Significant improvement of electrical distribution
  - Additional cooling equipment
- People have done (almost heroic) efforts to maintain the site in (near normal) operating conditions
- More on this during the visit to the facility later this afternoon

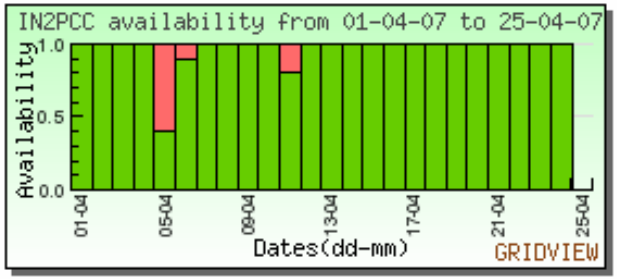
Site Availability

Site availability daily score: Oct 2006 – Mar 2007



IN2P3-CC **av.reliability last 3 mths** **77%**

Overall Service Availablity for site IN2PCC : Daily Report



Site availability daily score: April 2007

Sources: http://lcg.web.cern.ch/LCG/MB/availability/site_reliability.pdf http://gridview.cern.ch

Current work

- Top priority: integrating the grid services to the standard operations
  - Including on-call service
  - Monitoring & alerting, progressively documenting procedures, identifying roles and levels of service, etc.
  - Strong interaction with people developing the grid operations portal (CIC)
- Consolidation of the services
  - Deploy for availability
    - Hardware redundancy
    - Usage of (real or virtual) stand-by machines
- Assigning job priorities based on VOMS roles & groups
  - Interim solution in place
  - Site information system to be modified

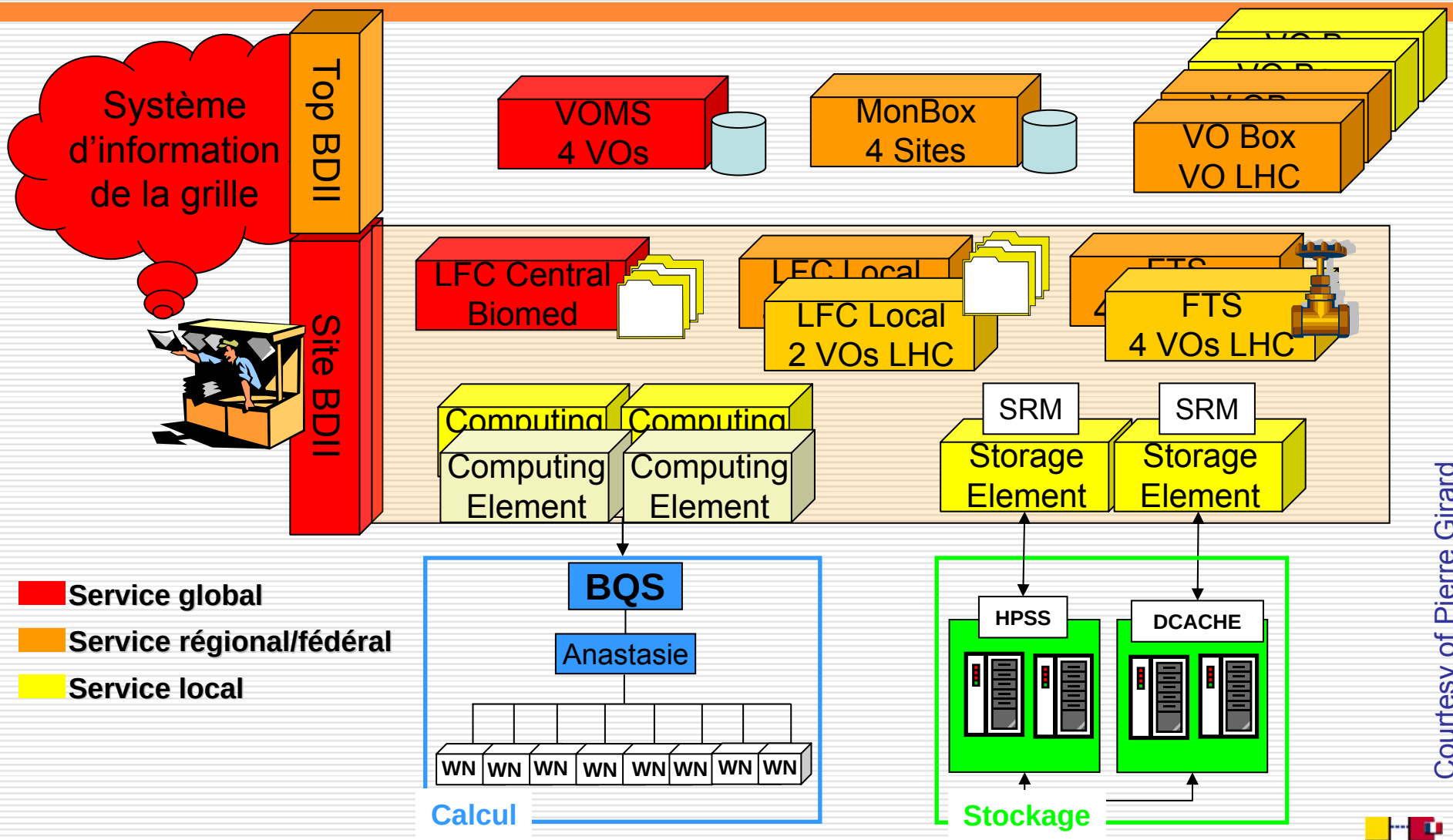
Current work (cont.)

- Continous development of BQS
  - for coping with the expected load in the years ahead
  - for making it more grid-aware
    - Keep grid attributes in the job records
      - *Submitter identity, grid name, VO name, etc.*
    - Scheduling based on grid-related attributes
      - *VOMS roles/groups, grid identity, etc.*
    - Allow/deny job execution based on grid identity
  - Development of gLiteCE and CREAM compatible BQS-backed computing element

Current work (cont.)

- Continous work for internal reconfiguration of HPSS according to the (known) needs of LHC experiments
- Trying to understand how the data will be accessed
  - What are the required rates for data transfers between MSS→disk→worker nodes and backwards…
  - ..for each one of the several kind of job (reconstruction, simulation, analysis, …)
- Job profiling
  - Studying the observed usage of memory for LHC jobs
    - Memory requirements have a significant impact on budget and on the capacity of our site to efficiently exploit the purchased hardware
- Understanding how to build the Analysis Facility
  - Ongoing discussion with Atlas and CMS

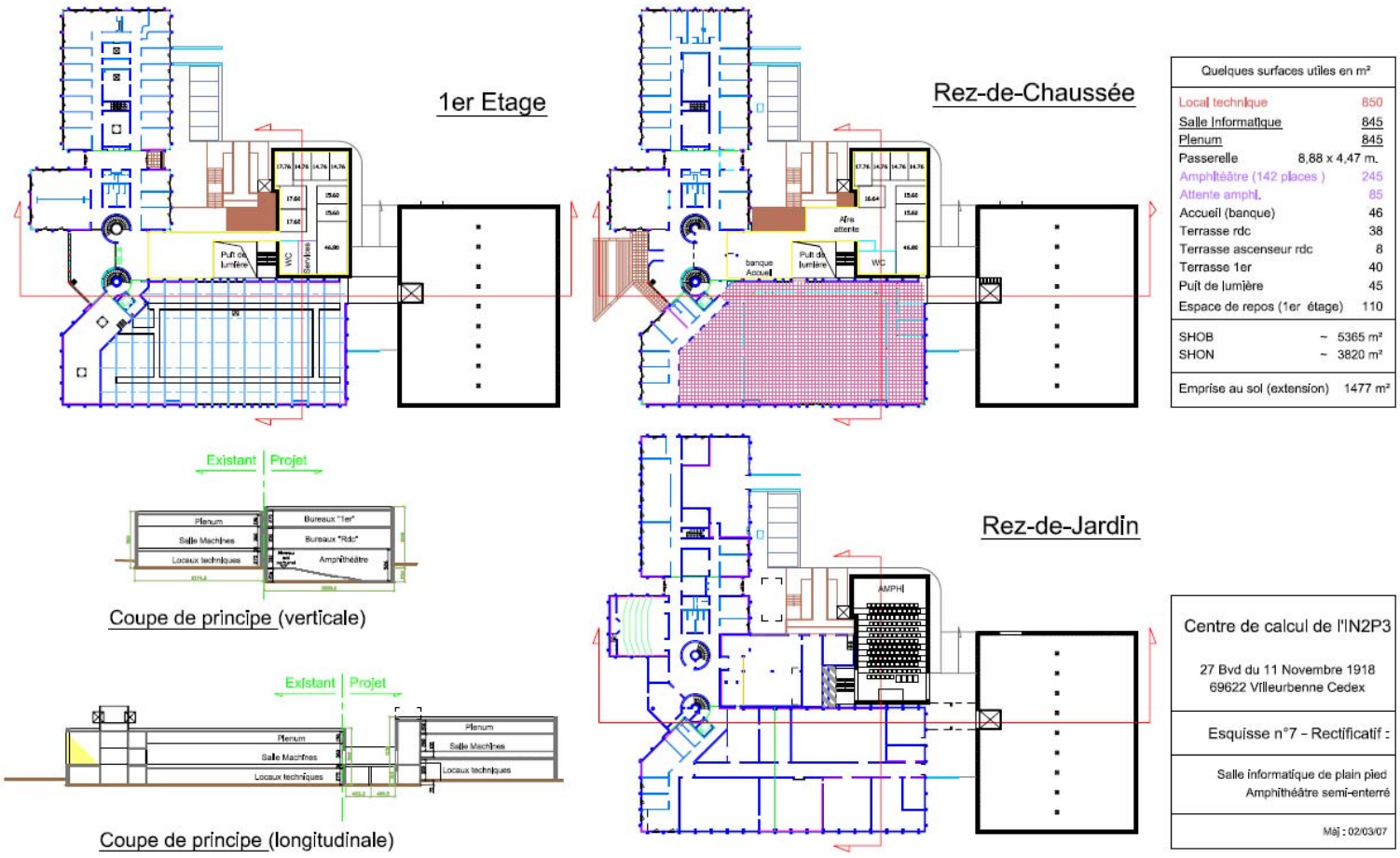
Overview of grid services
(by the end of June 2007)



New building

- On-going project for building an additional machine room
  - 800 m² floor space
  - Electric power for computing equipment: 1 MW at the beginning, with capacity for increasing up to 2,5 MW
- Offices: for around 30 additional people
- Meeting rooms, 140+ seats amphitheatre
- Target availability: mid 2009

New building (cont.)



What's next

- In the coming presentations you will find the detailed status and plans of the
  - Network infrastructure
  - Grid services
  - Storage infrastructure and data transfers
  - Databases
  - Site operations

Questions

Fabien Wernli, 2006